Norbert Lossau

Suchmaschinentechnologie und Digitale Bibliotheken – Bibliotheken müssen das wissenschaftliche Internet erschließen[1]

Foto Neue Westfälische

Norbert Lossau

Will Google, Yahoo and Microsoft become the only avenues for accessing the world of knowledge in the year 2010? The author of this article makes the case for concerted action on the part of libraries which would lead to the creation of dependable, high-quality search services for academic information for research and teaching via state-of-the-art search engine technology. The need for such a service has arisen from the numerical explosion of scientifically relevant documents that are often accessible only through the internet, but not (or not adequately) researchable through libraries' information portals. This article describes the possibilities for libraries and other information providers to cooperate in a national and international context in order to construct an open, shared academic internet index which can be used in any local environment in modular format. It is noted that a user-friendly design of the new system will need to take into account the current, general internet search machines, which have achieved a high level of popularity and user acceptance within quite a short period of time.

Werden Google, Yahoo und Microsoft die einzigen Zugänge zum weltweiten Wissen im Jahre 2010 darstellen? Der Autor setzt sich für eine konzertierte Aktion der Bibliotheken ein, um mittels »State-of-the-Art«-Suchmaschinentechnologie verlässliche, qualitativ hochwertige Suchdienste für wissenschaftliche Informationen in Forschung und Lehre aufzubauen. Die Notwendigkeit ergibt sich aus dem explosionsartigen Anwachsen wissenschaftlich relevanter Dokumente, die nicht selten ausschließlich über das Internet zugänglich sind und über derzeit verfügbare Informationsportale von Bibliotheken nicht oder nur sehr unzulänglich recherchiert werden können. Der Artikel beschreibt mögliche Wege der Kooperationen von Bibliotheken und anderen Informationsanbietern im nationalen und internationalen Kontext beim Aufbau eines offenen, verteilten wissenschaftlichen Internet-Indexes, der modular angelegt in beliebigen lokalen Umgebungen genutzt werden kann. Für die nutzerfreundliche Gestaltung des neuen Suchservices wird auf die Notwendigkeit der Berücksichtigung von etablierten, allgemeinen Internet-Suchdiensten eingegangen, die innerhalb kurzer Zeit eine hohe Popularität und Nutzerakzeptanz gewonnen haben.

Bibliotheksportale und externe Quellen

Die Suche nach relevanten Informationen über das World Wide Web ist inzwischen zu einem bedeutenden Wirtschaftsfaktor im globalen und kommerziellen Wettbewerb geworden. Der Markt wird beherrscht von weltweit operierenden Anbietern wie kommerziellen Internet-Suchmaschinen, »Informations-Portalen«, Verlagen und anderen Datenlieferanten.

Werden Google, Yahoo und Microsoft die einzigen Zugänge zum weltweiten Wissen im Jahre 2010 darstellen?

Wenn Bibliotheken in einem ihrer zentralen Dienstleistungsbereiche – dem Zugang zur Information – nicht bedeutungslos werden wollen, müssen sie die Herausforderungen annehmen, die mit der weltweiten Verfügbarkeit wissenschaftlicher Information und der Existenz und dem Wachstum des wissenschaftlichen Internets einhergehen.[2]

Wie definieren Bibliotheken wissenschaftlich relevanten (Online) Content?

Bibliotheken verstehen sich als zentrale Anbieter von Informationen für ihre Klientel in Universitäten und Forschungseinrichtungen. Doch wie definieren sie wissenschaftlichen Content? Ein Blick auf die Portale digitaler Bibliotheken zeigt, dass wissenschaftliche Internet-Quellen hier nur vereinzelt integriert werden. Man findet Bibliothekskataloge, elektronische Zeitschriften, manchmal auch E-Books – aber meist sind dies nur digitalisierte gedruckte Materialien, auf denen schon immer der Fokus in der Erwerbungspolitik der Bibliotheken lag. Auch der Zugang zu Datenbanken ist seit langem Bestandteil solcher digitaler Bibliotheken. Der Zugang zur Information selbst läuft über die etablierten Kanäle – Verlage, Buchhändler oder Zeitschriftenagenturen.

Die Digitalisierung des Publikationsprozesses und die Verbreitung des WWW führten zu einem starken Zuwachs von Inhalten in Typ und Formaten, u.a. digitalisierte Sammlungen, Websites von Fakultäten und Forschungseinrichtungen, Webserver für Konferenzen und Tagungen, Preprint/E-Print-Server und – in zunehmendem Maße – Speicher und Archive verschiedener Institutionen sowie eine große Anzahl E-Learning-Kurse und -Objekte. Diese Inhalte werden – wenn überhaupt – meist nur in separaten Linklisten oder Datenbanken angeboten, nicht aber als integraler Bestandteil eines Bibliotheksportals.

Warum Bibliotheken eher zurückhaltend sind, diese unbefriedigende Situation zu beseitigen, hat unterschiedliche Gründe: Da die Qualität einer Bibliothek noch immer an der Anzahl lokal verfügbarer Medien und Dienstleistungen gemessen wird, genießen diese grundsätzlich eine höhere Wertschätzung als solche Dienste, die nur über externe Zugänge zur Verfügung stehen. Bibliotheken verstehen sich nach wie vor eher als Sammlung lokal verfügbarer Quellen denn als Portal zum weltweiten Wissen.

Weitere Vorbehalte resultieren aus der Tatsache, dass es keine Garantie gibt, dass ein externer Anbieter auch auf lange Sicht Informationen bereitstellt. »Tote« Links waren schon immer ein potenzielles Problem für Portale zum Zugriff auf externe Quellen. Der langfristige Zugang zu Informationen ist jedoch einer

der Grundwerte, die Bibliotheken besitzen. Es müssen daher Wege gefunden werden, wie sichergestellt werden kann, dass auch externe Quellen einen solchen Langzeit-Zugang bieten.

Andere Ursachen mögen in der Tatsache begründet liegen, dass es schon immer einen natürlichen Widerstand gegenüber Veränderungen in bewährten Erwerbungspraktiken und Arbeitsabläufen gab. Auch die komplexe Kombination von Fachkenntnissen, technischem Know-how und traditionellem Erwerbungs-Know-how stellt eine gewisse Hürde dar.

Dieser neue Typ von »Erwerbung« – als eingeführter Arbeitsprozess in der Bibliothek – wird die Neuorganisation bestehender Strukturen erforderlich machen, mit möglichen Konsequenzen für Kosten und Ressourcen.

Sind sich die Bibliotheken des tatsächlichen Umfangs wissenschaftlich relevanter Online-Quellen bewusst?

Bibliotheken konzentrieren sich derzeit auf den Aufbau lokaler Digitaler Bibliotheken und die (gleichzeitige) Recherche in einer begrenzten Anzahl lizenzierter und frei zugänglicher Datenbanken. Man fragt sich, ob Bibliotheken der tatsächliche Umfang wissenschaftlicher Quellen bewusst ist, die bereits über das Internet zugänglich sind? Auch wenn es keine genauen Statistiken über die Größe des Web und die Anzahl der Internet-Seiten gibt, liefern einige Studien zumindest Annäherungswerte. Eine Untersuchung von Michael Bergman aus dem Jahre 2001 über das »Deep Web«[3] zeigt die Dimensionen auf: Bergman spricht von ca. 1 Mrd. frei zugänglicher Internet-Seiten im »visible web«[4] und fast 550 Mrd. Seiten im »deep web« oder »invisible web«. Den Zuwachs des »visible web« über die letzten Jahre kann man daran erkennen, dass Google die Größe seines Indexes (d. h. Seiten aus dem »visible web«) derzeit (Juni 2004) mit 4,2 Milliarden angibt (gegenüber 3,3 Mrd. Seiten im Jahre 2003).

Für Forschung und Lehre ist das »invisible web« von besonderem Interesse, da es zum überwiegenden Teil qualitativ hochwertige Inhalte in freien und lizenzierten Datenbanken, Primärdaten (z. B. meteorologische Daten, Finanzstatistiken, Quelldaten für Bioforschung usw.) oder die große und immer noch zunehmende Menge historischer Quellen umfasst, die digitalisiert werden. Auch in Bezug auf wissenschaftlich relevante Quellen gibt es keine verlässlichen Zahlen. Die wissenschaftliche Suchmaschine Scirus[5] liefert hier erste Anhaltspunkte. Im Mai 2004 hatte Scirus 167 Mio. wissenschaftliche Internet-Seiten indexiert, was ca. 4 % des Google-Indexes von 4,2 Mrd. Seiten entspricht. Auch wenn Scirus einige Quellen aus dem »invisible web« beinhaltet, stammt doch der größte Anteil aus frei zugänglichen Internet-Seiten mit wissenschaftlichem Bezug.[6] Wendet man den 4 %-Faktor von Scirus zu Google auf die Größe des »invisible web« im Jahre 2000 (550 Mrd. Web-Seiten) an, erhält man die erstaunliche Zahl von 22 Mrd. Seiten mit wissenschaftlich relevantem Inhalt.

Die Vision

gemeinsamer Suchindex von Bibliotheken als offener, verteilter Verbund

Was müssen die Bibliotheken ändern? Anstelle einer in hohem Grade zersplitterten Informationslandschaft, die den Nutzer zwingt, zahlreiche voneinander unabhängige Server zu nutzen, müssen Bibliotheken, gemeinsam mit anderen Partnern, *einen* Suchindex anbieten, der den Zugriff auf jede Art von wissenschaftlich relevanter Information ermöglicht. Bibliotheken tragen damit zu einem offenen, verteilten Verbund von Suchindexen bei, der eine Alternative zu den monolithischen Strukturen der kommerziell ausgerichteten Suchdienste bietet.[7]

Diese einzigartige Ressource stellt somit nicht irgendeine unbedeutende Untermenge eines kommerziellen Internet-Indexes dar, der von Werbung lebt, dessen Suchergebnisse nicht selten von der Werbeindustrie beeinflusst sind und der anderen Regeln hinsichtlich Relevanz und Nachhaltigkeit unterliegt. Die Bibliotheken müssen einen Rechercheservice anbieten, der qualitativ hochwertige Daten für Lehre und Forschung beinhaltet und in dem Vertrauenswürdigkeit und Langzeit-Verfügbarkeit des Suchindexes eine Hauptrolle spielen.

Bibliotheken zögern immer mehr, externe Portallösungen lokal zu integrieren, die alles »aus einem Guss«, mit eigener Oberfläche, anbieten. Solche »all-inclusive«-Lösungen bieten wenige Integrationsmöglichkeiten in das bestehende Bibliotheksangebot und haben zur Folge, dass nur ein weiterer Link auf eine externe Quelle im Informationsportal der Bibliothek erscheint. Suchservices der Zukunft sollten daher auf gemeinschaftlich erstellten, verteilten Datenquellen basieren und mit verschiedensten Such- und Browsing-Oberflächen ausgestattet sein, die sich nahtlos in die Oberfläche eines lokalen Informationsportals, einer virtuellen Fachbibliothek oder einer Lernumgebung einpassen lassen.[8] Bibliotheken, die diesen Suchindex nutzen, müssen in der Lage sein, nur die Datenquellen anbieten zu können, die speziell für ihre Klientel von Interesse sind. Die Such- und Browsing-Oberflächen und -Möglichkeiten müssen vollständig an Layout und Funktionalität der lokalen Web-Seiten angepasst werden können, um für den Nutzer ein einheitliches »look

nahtlose Einbindung in die Oberfläche lokaler Informationsportale

& feel« zu erreichen oder um fachspezifische Navigationselemente einbinden zu können.[9]

Dieser neue, wissenschaftliche Such-Index sollte die Vorzüge von Internet-Suchmaschinen – einfache Bedienung, schnelle Antwortzeiten – mit den Vorzügen einer klassischen Bibliothek – Relevanz und Qualität der Quellen – kombinieren.

Bereits existierende wissenschaftliche Portale

nationale und internationale Initiativen als Antwort auf Zersplitterung der Informationslandschaft

In den letzten Jahren ist die Zersplitterung der Informationslandschaft als gravierendes Problem erkannt worden und hat zu einer Reihe von Initiativen auf nationaler und internationaler Ebene geführt. Kürzlich wurde das »Vascoda«-Portal[10] gestartet, eine Kooperation zwischen Bibliotheken, Informationsanbietern und ihren Partnern auf internationaler Ebene – gefördert vom Bundesministerium für Bildung und Forschung und der Deutschen Forschungsgemeinschaft. Im Projekt »The Scholars Portal« – unter der Federführung des amerikanischen »Research Libraries«-Konsortiums – entstand eine Reihe von Portalen für Universitäten in den USA.[11] Weiterhin sind zu nennen: Das »Resource Discovery Network« (RDN) in Großbritannien,[12] das europäische RENARDUS-Projekt,[13] das Internet-SCOUT-Projekt aus den USA[14] und natürlich die »Virtuellen Fachbibliotheken« der Sondersammelgebietsbibliotheken. Dies sind nur einige Beispiele, in denen auf gemeinschaftlicher Basis metadatenbasierte Zugänge auf weltweit verteilte Quellen entstanden sind. Basierend auf der OAI-Initiative unterstützen Bibliotheken und ihre Service-Zentralen die entstehenden »OAI-Registries« als zentrale Zugangspunkte zu weltweit verteilten OAI-Repositories.[15]

Aus Sicht der Bibliotheken sind dies ohne Zweifel erfolgreiche Beispiele dafür, einen integrierten Zugang auf verteilte Quellen zu bieten. Dennoch sind weitere Entwicklungen erforderlich, will man wirkliche Service-Angebote für den Endnutzer aufbauen, die Wissenschaftler und Studierende zufrieden stellen. Anders als Internet-Suchmaschinen konzentrieren sich diese Projekte in erster Linie auf bibliographische Angaben, indem sie z. B. Informationen über die Quelle (einen Server oder eine Datenbank) geben und nicht die Recherche in der Quelle selbst (also in den Volltexten) ermöglichen. Natürlich können diese Angaben als Startpunkt in das »invisible web« genutzt werden. Ist in der Beschreibung eine URL genannt, kann diese dazu dienen, dass Internet-Suchmaschinen zumindest die Startadresse der Quelle indexieren.[16] Die eigentliche Arbeit für die Bibliotheken beginnt jedoch erst, nachdem eine Quelle lokalisiert worden ist, denn Standard-Internet-Suchmaschinen können Inhalte aus dem »invisible web« nicht ohne weiteres indexieren. In diesem »invisible web« gibt es eine unüberschaubare Anzahl verschiedener Datenformate und unterschiedlichste technische Implementierungen in Datenbanken und auf Servern.

Problem: Metadaten statt Volltexte

Der Einfluss von Internet-Suchmaschinen auf die Bibliotheken

Die Suche nach Informationen über das World Wide Web ist, wie bereits angesprochen, zu einem bedeutenden Wirtschaftsfaktor im globalen und kommerziellen Wettbewerb geworden. Bibliotheken sind nur ein Anbieter auf diesem Markt. Weitere Interessengruppen sind z. B. Verlage, kommerzielle Internet-Suchmaschinen und andere Datenlieferanten.

Es ist deshalb von besonderer Bedeutung, einen Blick auf die potenziellen Kunden und ihr Nutzerverhalten zu werfen. Das mag für Bibliotheken zunächst trivial erscheinen, da sie quasi von Berufs wegen die Auffassung vertreten, die Bedürfnisse ihrer Nutzer zu beachten oder besser gesagt, das, was sie für die Bedürfnisse halten. Doch die neue Wettbewerbssituation zwingt Bibliotheken dazu, ihre Dienste weitaus genauer aus Nutzersicht zu betrachten. Insbesondere die Universitätsbibliotheken haben es dabei mit einem sehr heterogenen Nutzerkreis mit ganz unterschiedlichen Bedürfnissen zu tun. Studierende im Grundstudium stellen andere Anforderungen an ihre Bibliothek als Wissenschaftler und andere Informationsspezialisten und haben ein entsprechend unterschiedliches Nutzerverhalten. Ein Student im Erstsemester wird in einem Bibliothekskatalog nichts anderes als eine Suchmaschine sehen und dementsprechende Recherchen betreiben, wie er es z. B. von Google gewöhnt ist, während Experten die Besonderheiten bei der Recherche nach wissenschaftlichen Quellen verinnerlicht haben. Vor der Einführung des Internets war diese Differenzierung für die Bibliotheken nur insoweit relevant, als für die Nutzung der Bibliotheksdienstleistungen (gedruckter Katalog, Online-Katalog, Nutzung digitaler Medien) verschiedene Schulungen und Einführungen nötig waren. Heute können Nutzer eine ganze Reihe von anderen Online-Diensten benutzen (Internet-Suchmaschinen, Portale, Bibliothekskataloge anderer Bibliotheken). Die Nutzer sind inzwischen durch Internet-Suchmaschinen wie Google in der Lage, ihre eigene Auswahl unter verschiedenen Such-Tools zu treffen und an Informationen zu gelangen, ohne zuvor Schulungen oder Tutorials besuchen zu müssen. Während die Bibliothekare sich über die mangelnde Qualität der Suchindexe von Internet-Suchmaschinen beklagen, sind die Nutzer von der einfachen Handhabung

der Internet-Suchmaschinen begeistert und würden diese am liebsten für jede Art von Informationsrecherche nutzen.

Wie eine Studie an der Universitätsbibliothek Bielefeld aus dem Jahre 2002 ergab,[17] nutzen Studenten immer noch sehr häufig den lokalen Bibliothekskatalog, einfach deshalb, weil die Quellen, die dort verzeichnet sind, nicht über Internet-Suchmaschinen gefunden werden können. Grundsätzlich würden sie aber eine Suchmaske à la Google bevorzugen. Immer dann, wenn eine Recherche in einer Internet-Suchmaschine Erfolg verspricht – insbesondere wenn es um Volltexte geht – werden Suchmaschinen weitaus häufiger genutzt als z. B. Fachdatenbanken, auf die die Bibliothek den Zugriff anbietet. In Datenbanken, E-Journals etc. wird nur gesucht, wenn der Nutzer über entsprechende Recherchekenntnisse verfügt oder einfach aus Gewohnheit diese Quellen anstelle von Internet-Suchmaschinen verwendet. Doch in einigen Jahren wird es eine neue Generation von Wissenschaftlern geben, die mit der Nutzung von Internet-Suchmaschinen bereits aufgewachsen ist.

Die Nutzer schätzen an Suchmaschinen sowohl die einfache Bedienung als auch die flexible und übersichtliche Art der Präsentation des Suchergebnisses. Die überlegene Performance und die enorme Größe der Suchmaschinen-Indexe werden ebenfalls als selbstverständlich angenommen.

Herausforderungen für die Suchsysteme heutiger Informationsportale

Unter der Federführung der Universitätsbibliothek Bielefeld entstand das zentrale Portal für wissenschaftliche Hochschulbibliotheken in Nordrhein-Westfalen, die Digitale Bibliothek NRW, die seit 2001 als regulärer Service läuft.[18] Basierend auf den Erfahrungen, die in einen erfolgreichen Service mündeten, wurde wenig später in Bielefeld eine Projektgruppe gegründet, die sich mit den aktuellen Entwicklungen im wissenschaftlichen Internet beschäftigte und dabei einige Mängel feststellte, die viele Digitale Bibliotheken aufweisen:

Dateiformate, Volltextsuche

Die meisten Systeme beschränken sich ausschließlich auf die Recherche in Metadaten (bibliographische Angaben, Abstract). Die gleichzeitige Recherche in Volltexten wurde erst kürzlich eingeführt und beschränkt sich in der Regel auf wenige Dateiformate (u. a. HTML- und Text-Dokumente)

Die Erfassung unterschiedlichster Content-Typen

Digitale Bibliotheken, die eine simultane Recherche in verschiedenen Quellen ermöglichen, decken derzeit ganz überwiegend nur Bibliothekskataloge und Datenbanken mit einigen Volltext-Quellen (z. B. E-Journals) ab. Frei zugängliche, wissenschaftliche Internet-Quellen werden häufig nicht in diese Informationsportale integriert. Wenn sie überhaupt erfasst werden, dann nicht selten nur als Link auf die Quelle in Form einer Linkliste oder innerhalb einer Datenbank, versehen mit referenzierenden, manchmal auch bewertenden Metadaten zu Online-Inhalten.

Wissenschaftler veröffentlichen ihre Preprints oder Postprints auf den Web-Seiten der Fakultät oder der Forschungseinrichtung. Auf den Web-Seiten von wissenschaftlichen Kongressen sind Präsentationen und Vorträge abgelegt. Auf großen, internationalen Preprint-Servern – oftmals von Wissenschaftlern organisiert – sind hunderttausende Dokumente gespeichert. Die Erstellung eigener E-Learning-Objekte nimmt immer mehr zu.

Preprint-Server, Hochschulschriftenserver, Digitalisierungsserver

Und die Bibliotheken? Auch sie tragen zur Zunahme von Online-Content bei. Seit fast 15 Jahren werden gedruckte Bestände systematisch digitalisiert. Mittlerweile gibt es hunderte, wenn nicht gar tausende Server mit digitalisierten Sammlungen. Meist sind die Systeme völlig unabhängig voneinander aufgesetzt. Die Aktivitäten in den Universitäten zum Aufbau eigener Hochschulschriften-Server haben gerade erst richtig begonnen. Auf lange Sicht ist es ihr Ziel, alle Dokumente, die an einer Universität produziert werden, auch auf eigenen Dokumenten-Servern abzuspeichern. Während der Aufbau solcher Datenspeicher aus strategischen Gründen erwünscht ist (Stichworte »Open Access«, Langzeit-Verfügbarkeit), erfordert die dadurch steigende Zahl von Online-Servern noch mehr Anstrengungen auf der Recherche-Seite.

Begrenzte Skalierbarkeit und Performance

Die Mehrheit der Portale funktioniert nach dem Prinzip der Metasuche. Eine Anfrage wird abgeschickt und in die Suchsprachen der Zielsysteme umgewandelt. Die Resultate aus den verschiedenen Quellen werden sequenziell zusammengeführt und in einer gemeinsamen Trefferliste ausgegeben. Die Probleme, die durch dieses Metasuche-Prinzip entstehen, sind offensichtlich: Durch die sequenzielle Abarbeitung und insbesondere durch die Abhängigkeit von den Zielsystemen in punkto Geschwindigkeit und Suchmöglichkeiten ist die Skalierbarkeit dieser Systeme begrenzt. Die Performance nimmt ab, je mehr Quellen eingebunden und durchsucht werden.

Suchkomfort

Alle Suchsysteme folgen den Prinzipien der Booleschen Logik, die auf den Recherchekomfort großen Einfluss hat. Internet-Suchmaschinen integrieren – basierend auf linguistischen Analysen, Wörterbüchern und Lexika – mehr und mehr Funktionen der »approximativen« Suche und ermöglichen dadurch eine größere Fehlertoleranz bei den ausgewählten Suchbegriffen. Im Gegensatz dazu erfordert die ausschließliche Verwendung der Booleschen Logik beim Nutzer bereits Erfahrungen hinsichtlich der Auswahl der »richtigen« Suchbegriffe.

Recherchekomfort durch »approximative Suche«

Trefferlisten, Ranking (Relevanzbewertung)

Die Recherchesysteme digitaler Bibliotheken sind auf das Vorhandensein von bibliographischen Angaben (Autor, Titel, Erscheinungsjahr etc.) angewiesen, um eine sortierte Trefferliste ausgeben zu können. Diese Sortierung folgt im Prinzip immer noch der Sortierung in traditionellen Zettelkatalogen und ist sehr unflexibel. Internet-Suchmaschinen haben Ranking-Methoden entwickelt, die sowohl auf statischer als auch auf dynamischer Echtzeit-Analyse des Suchergebnisses basieren und die es dem Nutzer erlauben, seine Suchergebnisse nach verschiedenen Ranking-Kriterien zu sortieren. Dies ist ein sehr beliebtes Feature von Internet-Suchmaschinen.

Mängel der Internet-Suchmaschinen für den Einsatz in wissenschaftlichen Bibliotheken

Unter den Bibliotheken hat die Diskussion über die Erschließung des wissenschaftlichen Internets gerade erst begonnen. Einige Institutionen haben sich entschlossen, einen Teil ihrer »unsichtbaren« Inhalte für Internet-Suchmaschinen freizugeben (z. B. bibliographische Datensätze aus Bibliothekskatalogen oder E-Prints aus Hochschuleinrichtungen) – ein sehr pragmatischer und kostengünstiger Weg, um qualitativ hochwertige Inhalte »sichtbar« und damit auch über Internet-Suchmaschinen auffindbar zu machen. Die Gründe, bereits existierende Internet-Services zu benutzen, sind einleuchtend. Bibliotheken sollten aber zusätzlich eigene Strategien und Konzepte entwickeln, um das wissenschaftliche Internet zu erschließen.

Probleme: Ranking-Methoden, fehlende Langzeitverfügbarkeit

In erster Linie sind Dienste wie Google rein kommerziell ausgerichtet. »A search engine's primary business is to obtain revenue through advertising« – lautet das Fazit in einem Artikel von Rita Vine, zitiert im D-Lib Magazine.[19] Das Ranking der Suchergebnisse in Suchmaschinen erfolgt nach verschiedenen Methoden, eine davon ist das Ranking nach Bezahlung. Wer am meisten bezahlt, um bei einem bestimmten Suchbegriff aufgelistet zu werden, landet ganz oben in der Trefferliste. Dies ist ebenso ein Nachteil wie die fehlende Garantie der Langzeitverfügbarkeit eines Suchmaschinen-Indexes, d. h. dass der Index einer Woche auch noch alle Quellen der letzten Woche enthält. Die Business-Konzepte werden von den Suchmaschinenbetreibern streng geheim gehalten. Die kürzlich erfolgte Übernahme der Suchmaschine Alltheweb durch Yahoo – einschließlich der Ersetzung des bestehenden Indexes durch den Yahoo-Index – zeigt die Unsicherheiten, denen kommerzielle Suchmaschinen unterliegen: Der neue Yahoo-Index scheint deutlich kleiner zu sein als der Alltheweb-Index, und er bietet weniger Recherchemöglichkeiten als Alltheweb. Ein weiteres Beispiel ist das »Google Directory«, ein Abzug aus dem »Open Directory Projekt«, dem weltweit größten Internet-Verzeichnis. Auf der US-Homepage von Google (google.com) wurde der Link auf das »Google Directory« durch die Shop-Suchmaschine »Froogle« ersetzt und verschwand damit von der Homepage von google.com. Auch auf den anderen englischsprachigen »Ablegern« von Google (google.ca, google.co.uk) verschwand der Link auf das »Directory« (das Verzeichnis wird zwar weiterhin angeboten, ist aber praktisch aus Sicht der Nutzer verschwunden, da es nicht mehr von der Google-Startseite verlinkt wird). Interessanterweise hat diese Veränderung auf den nicht-englischsprachigen Google-Seiten (z. B. google.de) noch nicht stattgefunden.

Um Missverständnissen vorzubeugen: Kommerzielle Suchmaschinenbetreiber haben selbstverständlich das Recht, jegliche Art von Service anzubieten, der für sie profitabel ist. Bibliotheken sollten jedoch nicht annehmen, dass diese Services bibliothekarischen Kriterien der intellektuellen Qualitätsbewertung und Verlässlichkeit bei der Suche folgen.

Suchmaschinenbetreiber konzentrieren sich auf Inhalte, die automatisch indexiert werden können. Die manuelle Konvertierung von Daten oder die zeitaufwändige Analyse von Dateiformaten und Übertragungsprotokollen ist für kommerzielle Anbieter nicht von Interesse, da es bei ihnen in erster Linie darauf ankommt, Woche für Woche einen noch größeren Index anbieten zu können. Wissenschaftliche Inhalte sind – wie schon beschrieben – oftmals Teil des »invisible web« und können daher von Internet-Suchmaschinen nicht automatisiert indexiert werden. Institutionen, die ihre Daten in den Google-Index bringen wollen, müssen sich darüber im Klaren sein, dass dies Konvertierungsarbeit für die Bibliothek bedeuten kann und dass die Daten häufig nicht über Standard-Schnittstellen wie OAI (inkl. Dublin-Core-Felder) kon-

vertiert werden können, wie es eigentlich von Bibliotheken erwünscht ist.

Bibliotheken haben schon oft ihre Bedenken über die Qualität der Indexe von Internet-Suchmaschinen und die Vermischung von wissenschaftlich hochrelevanten und ungeprüften Inhalten zum Ausdruck gebracht. Sie vermissen häufig die Zuverlässigkeit der Quellen, wie sie z. B. bei den oben beschriebenen Recherche-Angeboten von Digitalen Bibliotheken usw. vorhanden sind.

Nicht zuletzt ist die Hard- und Software-Architektur kommerzieller Suchmaschinenbetreiber sehr kostenintensiv. Experten sprechen von Kosten in Millionenhöhe, um den Index einer großen Suchmaschine einmal pro Woche zu aktualisieren.[20] Daher ist keine Bibliothek in der Lage, einen solchen Service allein anbieten zu können. Bibliotheken müssen also Verbünde bilden und kooperieren, um ein gemeinsames Netzwerk aufzubauen.

Neben den offensichtlichen Mängeln von Internet-Suchmaschinen, die sich Bibliotheken vergegenwärtigen müssen, sollten hier auch noch einige Mythen über Suchmaschinentechnologie aufgeführt werden: Die Aussage »Suchmaschinentechnologie eignet sich nicht zur Indexierung strukturierter, hochqualitativer Daten« ist genauso falsch wie die Aussage »Suchmaschinentechnologie – das sind doch nur einzeilige Suchmasken«. Bei Aussagen wie diesen werden Anwendungen der Technologie mit dem tatsächlichen Potenzial der Technologie vermischt.

Kommerzielle Internet-Suchmaschinen und Suchmaschinentechnologie

Mehrheitlich richten sich die Vorbehalte der Bibliothekare nicht gegen die Technologie selbst, die hinter dem Index liegt, sondern gegen das Konzept kommerziell ausgerichteter Suchmaschinen. Doch warum sollten Bibliothekare nicht hinter dieses Konzept blicken und sich auf die eigentliche Suchmaschinentechnologie fokussieren?

Unter den großen Technologieanbietern, die keinen eigenen kommerziellen Suchindex anbieten, findet sich eine Firma, die regelmäßig als Nr. 1 gelistet ist: Die norwegische Firma Fast Search & Transfer (FAST), entstanden als Ausgründung der »Norwegian University of Science and Technology« (NTNU),[21] ist einer der Marktführer auf diesem Sektor und für technische Innovationen wiederholt ausgezeichnet worden. Praktische Anwendungen der eingesetzten Suchmaschinentechnologie für Internet-Suchmaschinen wie Alltheweb haben gezeigt, dass sie für den Einsatz mit riesigen Datenmengen geeignet ist, auch wenn FAST selbst die Technologie explizit als »Enterprise Search Solution« bezeichnet, um zu betonen, dass die Technologie über reine Internet-Suchmaschinen hinaus ihr Einsatzgebiet findet.[22]

Nach einer Evaluierungsphase[23] hat sich die Universitätsbibliothek Bielefeld entschlossen, FAST als Such-Technologie einzusetzen, um die Vorzüge und Verwendbarkeit für digitale Bibliotheken zu testen. Dabei sollte das Rad nicht neu erfunden und eine ganz neue Suchmaschinentechnologie entwickelt werden, sondern es sollten neue Funktionalitäten und Module zu einem bestehenden Produkt hinzugefügt werden: z. B. Datenkonnektoren (für OAI-Daten und Datenbanken), neue Such- und Browsing-Oberflächen oder verbessertes Ranking und Ergebnispräsentation für wissenschaftliche Inhalte. Einige erfreuliche Aspekte im Test mit FAST waren: Eine umfangreiche Dokumentation, ein breites Angebot leistungsfähiger Module (z. B. linguistische Methoden, approximative Suche), leistungsfähige Programmierschnittstellen (APIs) und die Unterstützung verschiedener Standards (der Index basiert auf XML). FAST ist keine schlüsselfertige Lösung, sondern ein System, welches sich hervorragend an eigene Bedürfnisse anpassen lässt. FAST wurde gewählt, um ganz pragmatisch mit einem Partner zusammenzuarbeiten, der innovationsorientiert ist und von Beginn an eine stabile und skalierbare Kerntechnologie bietet. Alternativen, insbesondere aus dem Open-Source-Bereich, wie z. B. »Nutch« / »Jakarta Lucene«,[24] werden aber ebenfalls beobachtet und die Möglichkeiten für eine Zusammenarbeit ausgelotet.

kostenintensive Hard- und Software-Architektur kommerzieller Suchmaschinenbetreiber

Zusätzliche Anforderungen an Recherche-Services im wissenschaftlichen Bereich

Die Diskussion in der Universitätsbibliothek Bielefeld – in Zusammenarbeit mit Kollegen aus anderen Bibliotheken – mündete in eine Liste von Anforderungen, die, zusätzlich zu den Standardfunktionalitäten der heutigen Suchmaschinentechnologie, für den Einsatz im wissenschaftlichen Umfeld unerlässlich sind:

Indexierung ausschließlich geeigneter Quellen

Es wurde schon angesprochen, dass bereits existierende Datenbanken und Verzeichnisse im Bibliotheksbereich (z. B. die virtuellen Fachbibliotheken, Portale auf regionaler und nationaler Ebene) eine zuverlässige Basis für intellektuell erschlossene Quellen bieten. Bibliotheken, die Suchmaschinentechnologie einsetzen, könnten diese Quellen indexieren. Diese Vorgehensweise sichert dem neuen wissenschaftlichen Such-Index eine gewisse Objektivität, die den meisten kommerziellen Internet-Indexen fehlt.

Einsatz der Such-Technologie FAST an der UB Bielefeld

Umgang mit heterogenen Daten

Da die Nutzer mit verschiedenen Datenformaten und Quellen konfrontiert werden, erfordern Recherche- und Navigationsoberflächen besondere Aufmerksamkeit bei der Evaluation und Entwicklung. Metadaten, Volltexte, Bilder und Multimedia sind einige der bekanntesten Beispiele für die Heterogenität von Quellen, die in einem virtuellen Index zusammengefasst werden können. Die intelligente Aufbereitung und Kennzeichnung verschiedener Quellen für Recherche, Such-Verfeinerungen und Navigation ist eine wesentliche Anforderung an das integrierte Angebot heterogener Daten.

Ziel: flexible Systemarchitektur

Erweiterte Browsing-Funktionalitäten

Der Begriff »Suchdienst für das wissenschaftliche Internet« beinhaltet auch die Navigation zur gesuchten Information. Als etablierte und z.T. bereits realisierte Tools stehen wissenschaftliche Klassifikationen, fachspezifische Thesauri und Cross-Konkordanzen zur Verfügung, die eine fachübergreifende Suche ermöglichen. Entwicklungen, die aus dem Forschungsbereich des »semantic web« entstehen, sollten auch für die Suche im wissenschaftlichen Internet berücksichtigt werden.

Flexible Relevanzbewertung und Sortiermöglichkeiten für die Trefferliste

Die Trefferanzeige des neuen Such-Services sollte so beschaffen sein, dass sie verschiedene Sichten auf eine spezifische Trefferliste ermöglicht. Relevanzbewertung und Sortiermöglichkeiten können sowohl »statischen« Kriterien folgen (Sortierung nach Autor, Titel, Jahr, Klassifikation u.a.) als auch dynamisch generiert werden (z.B. durch Echtzeitanalyse der Dokumente in Bezug auf Semantik oder Syntax).

Digitale Bibliothek NRW

Automatische Extraktion von Metadaten

Metadaten haben eine hohe Priorität für das Suchen und Browsen in wissenschaftlichen Quellen. Da jedoch ein beträchtlicher Anteil wissenschaftlicher Quellen über keinerlei Metadaten verfügt (z.B. viele Dokumente von Wissenschaftlern auf Fakultäts-Servern), wäre es sehr sinnvoll, Dokumenten-Analyse-Werkzeuge zu entwickeln, die zumindest einen Grundbestand bibliographischer Informationen aus den Volltexten automatisch generieren könnten.[25]

Die Architektur von Suchdiensten der nächsten Generation

Systemmodularität

Heutige Suchdienste sind zumeist starr in digitale Bibliothekssysteme eingebunden. Eine unflexible Systemarchitektur entspricht nicht mehr dem Stand der Technik. Anbieter integrierter Bibliothekssysteme haben auf diese Entwicklung reagiert und bieten zunehmend separate lokale und zentrale Module (z.B. für die Verbundkatalogisierung) an. Neue Anforderungen an die Bibliotheken wurden bisher durch das Aufsetzen neuer Systeme realisiert, z.B. digitale Bibliotheken, digitale Sammlungen oder E-Print-Server. Wenn wir einen Blick über das Bibliotheksumfeld hinaus in die Universitäten werfen, erweitert sich die Palette zusätzlicher Systeme noch einmal erheblich. Die steigende Zahl unterschiedlicher Systeme im Zusammenhang mit den gestiegenen Kosten für Anschaffung und Pflege dieser Systeme hat in den Bibliotheken und auch auf Hochschulebene zu einem Diskussionsprozess darüber geführt, wie diese Systeme untereinander kommunizieren und wie einzelne Services in verschiedenen Systemen eingebunden werden können. Ein bekanntes Beispiel ist die Administration von Nutzerdaten und die Diskussion über ein »single sign-on« (das einmalige Einloggen in ein System, um die Dienste verschiedener Einrichtungen, die eine Authentifizierung verlangen, nutzen zu können). Auch neue Bereiche wie z.B. E-Learning erfordern eine Wiederverwendbarkeit bereits existierender Dienste in den Lernumgebungen.

Systemmodularität ist auch auf der Ebene digitaler Bibliotheken und Wissenschaftsportale sinnvoll. Ein Portal beinhaltet den Zugang auf verteilte, heterogene Quellen, angereichert um weitere Dienste. Die »Digitale Bibliothek NRW« bietet z.B. eine Metasuche über 60 Datenbanken und Kataloge sowie, als zusätzlichen Dienst, die Möglichkeit zur Prüfung der lokalen Verfügbarkeit von Quellen (als Print-Fassung in der Bibliothek vor Ort oder über Online-Zugang) oder zur Einrichtung von persönlichen Suchprofilen. Mittelfristig können flexible Systeme innerhalb bereits vorhandener digitaler Bibliotheken oder anderer Informationsportale eingesetzt werden, um die derzeit im Einsatz befindliche Suchtechnik durch eine moderne Suchtechnologie zu ersetzen, die den Stand der Technik widerspiegelt.

Interoperabilität von Diensten

Die Technologie der »web services« ermöglicht echte Interoperabilität zwischen Diensten und Systemmo-

dulen in unterschiedlichen Umgebungen. Sie arbeiten plattformunabhängig, sind einfach zu implementieren und eröffnen die Möglichkeit, externe Dienste nahtlos in die lokale Umgebung zu integrieren. An der Universität Bielefeld arbeitete die Bibliothek erfolgreich mit der EDV der Zentralverwaltung an der Realisierung einer »web service-Schnittstelle« für die zentrale Benutzerverwaltung und den Bibliothekskatalog. Der modulare Aufbau dieser Systeme ermöglicht und fördert eine ganzheitliche Sicht auf die Systemarchitektur der Universitäten. Es ist ein viel versprechender Trend, die derzeit an den Universitäten noch herrschende Trennung der vielen kleineren und größeren Anwendungen zu überwinden.

Das Konzept eines offenen, verteilten Verbundes für einen wissenschaftlichen Internet-Index – wie im Folgenden beschrieben – greift diese Entwicklungen auf. Dabei wird die Implementierung von »web services« vorgeschlagen, um den Einsatz eines vollständig kompatiblen Suchindexes und eine nahtlose Integration in die lokalen, regionalen oder fachlichen Informationsinfrastrukturen zu ermöglichen.

Aufbau eines offenen, verteilten Verbunds für einen wissenschaftlichen Internet-Index: Eine Aufgabe für Bibliotheken?

Durch die kontinuierlich und exponentiell wachsende Zahl von Online Content ist es unrealistisch daran zu glauben, dass *eine* Bibliothek alleine *einen* riesigen Web-Index aufbauen könnte, der alle Quellen beinhaltet. Selbst nur ein Teilangebot, z.B. die Indexierung wissenschaftlicher Online-Quellen eines Landes, wäre eine kaum zu bewältigende Aufgabe für eine einzige Institution. Deshalb sind Kooperationen zwischen den Bibliotheken erforderlich – und Bibliotheken haben damit bereits gute Erfahrungen, auch im internationalen Kontext. Wie kann so eine Zusammenarbeit in der Praxis aussehen?

Ist die Technologie für einen verteilten Internet-Index verfügbar?

Kommerzielle Internet-Indexe haben üblicherweise eine monolithische Architektur.[26] Die Technologie selbst erlaubt unter gewissen Gesichtspunkten bereits einen verteilten Index. Die Software von FAST ermöglicht den Aufbau eines virtuellen Master-Indexes, der aus verteilten »Child-Indexen« zusammengesetzt ist. Dies würde die Nutzung der FAST-Technologie auf allen Seiten erfordern. In einigen Business-Lösungen, die mit FAST-Software realisiert wurden, sind auch bereits externe, nicht FAST-basierte Internet-Suchmaschinen-Indexe eingebunden. Trotz einiger Funktionalitäten, die die verteilte Indexierung unterstützen, bedarf es hier weiterer Forschung.

Die derzeitigen Initiativen im Bereich der »Grid«-Technologie (die sich in einem weiteren Kontext auch mit dezentralisierten Massenspeichern befassen) könnten nützliche Technologien für den Aufbau eines verteilten Daten- und Zugangsverbundes liefern. Bibliotheken müssen die weiteren Entwicklungen in diesem Bereich sehr genau beobachten und offen für mögliche Kooperationen sein.

Implementierung von »web services«

Mögliche Rollenverteilung in einem Verbund für einen wissenschaftlichen Internet-Index

Es gibt unterschiedliche Rollen, die die Partner – abhängig von Finanzierungsmöglichkeiten, Personal und strategischen Überlegungen – prinzipiell übernehmen können:

Der »Nutzer«: Nahtlose Integration von Such- und Browsing-Oberflächen in lokale Umgebungen

Der leichteste Weg, um an diesem Verbund zu partizipieren, ist, den Index einfach zu übernehmen und unter der Forschergemeinde vor Ort Interesse an der Nutzung zu wecken. Dieses Konzept ist durch Google und andere kommerzielle Suchdienste populär geworden. Dabei wird eine einfache Suchmaske in die lokale Website eingebunden. Die abgeschickten Recherchen werden direkt an den Index der Suchmaschine gesendet und das Ergebnis wird als Standard-Trefferliste mit Layout und Ranking-Methoden des Suchdienstes geliefert.

Kooperation der Bibliotheken beim Aufbau des Web-Index

Der neue Verbund wird dieselben Funktionalitäten bereitstellen. Hierfür sind auf Anwenderseite keine Software-Investitionen[27] und nur minimaler Personaleinsatz[28] nötig.

Die nötigen Anpassungen, um Wünsche und Gegebenheiten vor Ort zu berücksichtigen, z.B. die voreingestellte Einschränkung auf bestimmte Quellen (nur aus bestimmten Fächern) oder die Realisierung verschiedener Such- und Browsing-Möglichkeiten (z.B. die Nutzung bestimmter Klassifikationen), müssen auf Server-Seite von einem der Verbundpartner realisiert werden, die den entsprechenden Teil des Indexes betreuen. Sind dafür individuelle Anpassungen nötig (Programmierung etc.), muss dies von der jeweiligen Institution finanziell erstattet werden.

Die »Datenlieferanten«: Bereitstellung von Online-Quellen für die »Index-Ersteller«

Über die reine Nutzung des Suchdienstes hinausgehend, werden alle Datenlieferanten (und insbesondere Bibliotheken) dazu ermutigt, ihre Daten, die momen-

tan noch zum »invisible web« gehören, den Index-Erstellern zur Verfügung zu stellen.

fünf Arbeitsschritte von der Auswahl der Quelle bis zur Bereitstellung des Zugangs

Die Universitätsbibliothek Bielefeld wurde während des letzten Jahres, in dem verschiedene Suchmaschinentechnologien getestet und Testumgebungen aufgesetzt wurden, von einer Reihe von Bibliotheken in Deutschland, Großbritannien und den USA durch die kostenlose Bereitstellung ihrer Daten unterstützt. Daten von mehr als zehn verschiedenen Institutionen wurden gespeichert, aufbereitet (»pre-processed«) und indexiert (u.a. von der Fakultät für Mathematik der Universität Bielefeld, der Niedersächsischen Staats- und Universitätsbibliothek Göttingen, der Technischen Informationsbibliothek/ Universitätsbibliothek Hannover, den Oxford University Library Services, den Universitätsbibliotheken in Michigan und Cornell, dem Springer-Verlag, dem Anbieter des »Zentralblatts für Mathematik« und der Universitätsbibliothek Bielefeld); eine Reihe verschiedener Quellen mit einer großen Anzahl verschiedener Formate, u.a. digitalisierte Sammlungen, Preprint-Server, elektronische Zeitschriften, die Archive verschiedener Institutionen, Bibliothekskataloge und Datenbanken. Eine Methode, um die Daten zu laden, war der Einsatz von »OAI-Harvestern«, bei denen – wie schon beschrieben – eine Reihe von Problemen zutage trat. Entscheidend für alle Partner war, dass sie ihre Daten nur für die Datensammlungs-Tools (»web crawler«, FTP, Metadaten-Harvester u.a.) zur Verfügung stellen mussten – die Quellen selbst blieben auf den Servern der Partner und damit auch unter ihrer Kontrolle. Dies hat sich sowohl für kommerzielle Partner als auch für Bibliotheken und andere Institutionen als besonders relevant erwiesen. Insbesondere die kommerziellen Partner haben erkannt, dass dieser neue Suchdienst einer Bibliothek in keiner Weise ihre bestehenden Lizenzvereinbarungen berührt, sondern nur eine weitere Zugangsmöglichkeit auf ihre Daten schafft. Der neue wissenschaftliche Internet-Index offeriert, analog zu bereits bestehenden Suchmaschinen-Indexen, eine neue, komfortable und qualitativ hochwertige Such- und Browsing-Plattform zu verteilten Quellen. Der Zugang zu den Quellen und deren Nutzung selbst bleibt unter der Kontrolle der Datenlieferanten. Ein wichtiger, zusätzlicher Aspekt war, dass die Identität der einzelnen Partner und Kollektionen in den Trefferlisten durch den Hinweis »Datenlieferant: Institution XYZ« klar markiert wurde.

konsequentere Nutzung erfolgversprechender Standards

Die »Index-Ersteller und Hosts«: Aufbau lokaler, regionaler oder fachspezifischer Indexe und die gemeinsame Nutzung innerhalb des Index-Verbundes

Bibliotheken und bibliothekarische Servicezentren, die selbst Kompetenzen im Einsatz von Suchmaschinentechnologie erwerben wollen, können ihren eigenen Index aufbauen und betreuen. Die Institutionen müssen eine entsprechende technische Infrastruktur inkl. Hardware,[29] Suchmaschinentechnologie und Personal bereitstellen, um den Datenworkflow bewältigen zu können. Die Universitätsbibliothek Bielefeld hat fünf Arbeitsschritte definiert, die nötig sind, nachdem eine Quelle ausgewählt wurde: Laden der Daten, Aufbereitung der Daten (»pre-processing«), Verarbeitung der Daten, Indexierung der Daten und Bereitstellung des Zugangs. Das Personal sollte im Umgang mit »Harvesting-Tools« (z.B. OAI), Datenbankkonnektoren und -protokollen (Z39.50, SRW, generische DB-Konnektoren), Perl und XSLT (für das »pre-processing«), Python (Verarbeitung der Daten), spezifischer Suchmaschinentechnologie (Indexierung der Daten) und schließlich noch mit PHP (Zugang und Navigation) vertraut sein.

Es gibt eine Reihe verschiedener Kriterien, falls Bibliotheken ihren eigenen Index aufbauen wollen. Die Bedürfnisse vor Ort haben natürlich höchste Priorität. Wenn die Bibliotheken aber an gemeinsamen, auch internationalen Aktivitäten teilnehmen möchten, ist ein Aktionsplan unerlässlich, um unnötige Doppelarbeit zu vermeiden.[30] Bibliotheken könnten nach verschiedenen Kriterien ihre Datenquellen auswählen: geografische Kriterien, bestimmte Fachgebiete oder ganz einfach pragmatische Überlegungen.

Zu den möglichen Partnern gehören auch alle Projekte und Initiativen, die weiter oben unter den »bereits existierenden wissenschaftlichen Portalen« genannt wurden. Organisation und Kommunikation ist für den Erfolg dieser Initiative von entscheidender Bedeutung.

Die Partner für den »technischen Support«: Tools, die beim Aufbau eines qualitativ hochwertigen Suchdienstes helfen

Es gibt viele Gebiete, bei denen Arbeitsteilung auf technischer Ebene helfen würde, Synergie-Effekte zu erzeugen, Ressourcen zu sparen und den gesamten Prozess im Aufbau eines wissenschaftlichen Internet-Indexes zu beschleunigen.

Es muss noch viel getan werden, um eine größere Hilfestellung beim automatischen Laden und Indexieren der Daten zu erreichen. Einen wichtigen Anteil daran haben Anbieter aus dem OAI-Bereich, indem sie an einer weitaus konsequenteren Nutzung erfolgversprechender Standards arbeiten. Da selbst schon ein Standard wie OAI Probleme bei der Indexierung mit sich bringt, ist absehbar, welche Schwierigkeiten die Quellen verursachen werden, die heute noch nicht OAI-

kompatibel sind. Die Universitätsbibliothek Bielefeld hat, basierend auf ihren Erfahrungen bei der Integration verschiedener Quellen in die Metasuche der Digitalen Bibliothek NRW, bereits verschiedene Skripte und Indexierungsabläufe entwickelt, die nötig waren, um nur 15 verschiedene Datenquellen in den bestehenden Suchmaschinen-Index zu integrieren.

Auf Seiten der Recherche- und Browsing-Oberfläche ist die Integration von Taxonomien, Klassifikationen etc. ein Schlüsselpunkt, der in einem gemeinsamen Ansatz angegangen werden könnte.

Partner außerhalb des Bibliotheksbereichs

Einige kommerziell ausgerichtete Datenlieferanten, wie z. B. Elsevier oder Thomson, haben bereits ihre bisherige Suchmaschinentechnologie ersetzt und neue Indexe aufgebaut.[31] Es scheint logisch, solche bereits existierenden Indexe in einen gemeinschaftlich geführten Indexverbund einzubinden, wenn diese Anbieter den noch festzulegenden Geschäftsbedingungen zustimmen.

Schlussfolgerung und Ausblick

Werden Google, Yahoo und Microsoft die einzigen Zugänge zum weltweiten Wissen im Jahre 2010 sein? Dieser Aufsatz setzt sich für eine konzertierte Aktion der Bibliotheken ein, um mittels »State-of-the-Art«-Suchmaschinentechnologie verlässliche, qualitativ hochwertige Suchdienste für Forschung und Lehre aufzubauen. Diese Bemühungen sind nicht als Konkurrenz zu kommerziellen Suchdiensten zu sehen, sondern stehen für eine logische Weiterentwicklung der traditionellen bibliothekarischen Service-Angebote in einem Wissenschaftsumfeld, welches inzwischen die ganze Welt umfasst. Ein Verbund für einen wissenschaftlichen Internet-Index, der – wie schon skizziert – von Bibliotheken betrieben wird, kann die besonderen Bedürfnisse der Wissenschaftler am besten befriedigen, da er einen komfortablen, verlässlichen und integrierten Zugang zu qualitativ hochwertigen Inhalten bietet. Doch um solch einen neuen Service anbieten zu können, sind Bibliotheken gezwungen, über den Tellerrand ihrer derzeitigen Informations-Infrastruktur zu blicken und von etablierten Suchdiensten zu lernen. Diese sind durch den Einsatz innovativer Technologien und eine ganz auf den Nutzer abgestimmte Vorgehensweise so populär geworden.

Wie können Bibliotheken vorgehen?

Die Bibliotheken müssen die Relevanz eines neuen Aktionsplanes anerkennen, um ihre derzeitigen Recherche-Angebote verbessern zu können. Man hat den Eindruck, dass viele Bibliotheken zwar bereits »irgendwie« einen Bedarf sehen, aber es für viele schwierig ist, das Problem beim Namen zu nennen. Die derzeitigen pragmatischen Ansätze, wissenschaftliche Inhalte für kommerzielle Suchdienste zur Verfügung zu stellen, sollten nur als erster Schritt zu einem neuen Service gesehen werden, der von den Bibliotheken selbst angeboten wird.

Für die Universitätsbibliothek Bielefeld ist die weitere Entwicklung von Recherche-Angeboten ein elementarer Teil ihrer Rolle als zentraler Informationsanbieter innerhalb der Universität. Sie hat theoretische Überlegungen erfolgreich in die Tat umgesetzt: durch die praktische Implementierung von Suchmaschinentechnologie und die Arbeit mit »echten« Ressourcen. Die ersten öffentlich zugänglichen »Demonstratoren« ihrer lokalen Suchmaschinentechnologie (BASE – Bielefeld Academic Search Engine) sind seit Ende Juni zugänglich: Ein Prototyp für einen Suchdienst für die Mathematik und ein allgemeiner Demonstrator für digitale Sammlungen.[32] Details über die praktischen Erfahrungen mit der eingesetzten Suchmaschinentechnologie werden in einem Folgeartikel behandelt.[33]

BASE – Bielefeld Academic Search Engine

Die Erfahrungen und das Ergebnis dieser Arbeiten mündeten in einen gemeinsamen Antrag der Universitätsbibliothek Bielefeld und des Hochschulbibliothekszentrums des Landes Nordrhein-Westfalen an die Deutsche Forschungsgemeinschaft. Die Forschungsgemeinschaft und das Bundesministerium für Bildung und Forschung wollen die gemeinsame Initiative »Verteilter Dokumentenserver« (VDS), angestoßen durch die AG Verbundsysteme, fördern, die innerhalb von »Vascoda«, der Deutschen Digitalen Bibliothek, angesiedelt ist.

Initiative »Verteilter Dokumentenserver«

Das VDS-Projekt verfolgt drei Hauptziele: Authentifizierung, Aufbau eines Metadaten-Registers und die nahtlose Integration mittels Suchmaschinentechnologie. Die Universitätsbibliothek Bielefeld begrüßt jede Art der Zusammenarbeit und würde sich freuen, ihren Teil eines Kern-Indexes mit anderen Partnern aus dem Bibliotheksbereich zu teilen, um den vorgeschlagenen Verbund aufzubauen.

Über Deutschland hinaus haben bereits andere Bibliotheken und Institutionen, wie z. B. die Oxford University Library Services und die Norwegische Nationalbibliothek, ihr Interesse an der neuen Technologie bekundet. Eine Reihe von Institutionen innerhalb der amerikanischen »Digital Library Federation« hat sich interessiert gezeigt, an einer konzertierten Initiative mitzuarbeiten. Alle andere Bibliotheken, Institutionen und Datenanbieter werden ermutigt, kommende Aktivitäten zu unterstützen. Während die Zusammenarbeit der Universitätsbibliothek Bielefeld mit der Firma FAST sehr erfolgreich war, werden sich andere Insti-

tutionen vielleicht für Alternativsysteme entscheiden. Tatsächlich ist das einzig Entscheidende die Offenheit jedes Systems, um Kompatibilität zwischen den Systemen für den geplanten Such-Index-Verbund zu erreichen. Wenn das Konzept selbst realisiert werden kann, hängt die Auswahl des verwendeten Systems ohnehin von einem technischen Benchmarking ab, das auch bereits existierende und zukünftige Standard-Internet-Dienste einbeziehen muss.

nationale und internationale Kooperation

Es ist zu erwarten, dass andere Bibliotheken ihre eigene lokale Suchmaschinen-Infrastruktur aufbauen werden. Ein (informeller) Verbund oder ein Forum wird entstehen, das dem Erfahrungsaustausch und der gemeinsamen Nutzung von Inhalten und Tools dient. Die Notwendigkeit für ein Vorgehen auf internationaler Ebene war noch nie so offensichtlich wie in diesem Fall. Es kann davon ausgegangen werden, dass Förderorganisationen auch im internationalen Kontext die Relevanz der skizzierten Maßnahmen sehen und aktiv werden, um die Bibliotheken in die Lage zu versetzen, eine ihrer wichtigsten Aufgaben zu erfüllen: die reichhaltigen Informationen aus dem wissenschaftlichen Internet zu erschließen und zum Nutzen von Forschung und Lehre bereitzustellen.

[1] Der Artikel ist eine Übersetzung des englischen Originalartikels »Search Engine Technology and Digital Libraries: Libraries need to discover the academic internet.« In: D-Lib Magazine, June issue 2004. www.dlib.org/dlib/june04/lossau/06lossau.html; doi:10.1045/june2004.lossau. – Übersetzt durch Sebastian Wolf, UB Bielefeld.

[2] Stephen Arnold, ein bekannter »Information Consultant« aus den USA, hat erst kürzlich in seinem Aufsatz »Information boundaries and libraries« die Gefahr angesprochen, dass Bibliotheken marginalisiert werden könnten. www.arnoldit.com/articles/elib_Feb2004_final.doc

[3] Michael K. Bergman (im Auftrag der BrightPlanet Corporation) »The deep web: Surfacing hidden value«. In: The Journal of Electronic Publishing, Michigan University Press, July 2001. www.press.umich.edu/jep/07-01/bergman.html

[4] Bergman benutzt den Begriff »surface web«.

[5] www.scirus.com

[6] Die Mehrheit der Inhalte aus dem »deep web« kommt aus der Medline-Datenbank und Elseviers »Science Direct«-Datenbank. Zu den Websites, die gecrawled wurden, gehören Domains wie ».edu« (58,5 Mio. Seiten) und ».ac.uk« (6,8 Mio. Seiten): www.scirus.com/srsapp/aboutus/#range

[7] Partner können im Prinzip Bibliotheksservicezentren, Verlage oder andere kommerzielle Datenlieferanten sein.

[8] »Web services« scheinen eine viel versprechende Technologie zu sein, die eine modulare Integration von externen Services in eine lokale Umgebung ermöglicht.

[9] Durch die Benutzung von Navigations-Tools wie z. B. Thesauri, Cross-Konkordanzen oder Klassifikationen.

[10] www.vascoda.de

[11] www.arl.org/access/scholarsportal/

[12] www.rdn.ac.uk

[13] www.renardus.org

[14] http://scout.wisc.edu

[15] Drei Initiativen wurden auf dem »DLF Spring Forum 2004« präsentiert: University of Michigan Libraries: www.diglib.org/forums/Spring2004/hagedorn0404.htm; University of Illinois, Urbana-Champaign: www.diglib.org/forums/Spring2004/spring04bios.htm#habing; OCLC: www.diglib.org/forums/Spring2004/young0404.htm

[16] Die FAST-Software zum Beispiel beinhaltet einen »file traverser« und einige generische Datenbankkonnektoren, um Daten aus externen, heterogenen Quellen zu laden.

[17] 625 Studierende (im Grund- und Hauptstudium) beantworteten den Fragebogen innerhalb einer Woche.

[18] Die Entwicklung der Digitalen Bibliothek NRW wurde koordiniert vom früheren Direktor der Universitätsbibliothek Bielefeld, Dr. Karl Wilhelm Neubauer, und wurde unterstützt durch einen bedeutenden Zuschuss des Ministeriums für Wissenschaft und Forschung. Das Projekt startete 1998 und endete offiziell 2001 mit der Aufnahme des Regelbetriebs durch das Hochschulbibliothekszentrum in Köln (HBZ). Die Portaltechnologie (IPS-System) wurde von der Firma IHS nach den Spezifikationen der Projektplanungsgruppe entwickelt.

[19] Zitiert durch Bonita Wilson in: »Using the Internet for Searching«, D-Lib Magazine, March 2004, Vol. 10, no. 3

[20] Stephen Arnold liefert hierzu folgende Zahl: »$24 million per year to index one billion content sources« (24 Mio. $ pro Jahr, um eine Milliarde Seiten zu indexieren). In: The future of search. Proceedings of the 26th Online Information Conference 2002, S. 51

[21] Fast Search & Transfer hat seinen Sitz in Oslo (Norwegen) mit Filialen in verschiedenen europäischen Ländern, den USA und Japan (www.fastsearch.com). Nach dem Verkauf seiner Internet-Produkte (wie z. B. Alltheweb) an Overture im April 2003 hat sich FAST auf die innovative Entwicklung seiner Suchmaschinentechnologie durch das eigene Forschungslabor in Norwegen und durch Kooperationen mit der Universitäten in München, Cornell, Penn. State (beide USA) und Trondheim/Tromsø (Norwegen) konzentriert.

[22] Kunden kommen aus multinationalen Unternehmen, e-Government, e-Health usw.

[23] Februar – August 2002, Convera, Google, Mnogo, Lucene (2003)

[24] www.nutch.org, http://jakarta.apache.org/lucene/docs/index.html

[25] Die UB Bielefeld arbeitet mit dem externen Forschungsinstitut von FAST in München zusammen.

[26] Auch wenn die Daten selbst – aus Performance-Gründen – auf großen Computer-Clustern verteilt liegen.

[27] Es gibt keine lokale Client-Software, die lizenziert oder installiert werden muss.

[28] Zum Beispiel für die Einbindung von Templates in die eigene Website.

[29] Üblicherweise »low-cost«-Hardware, wie z. B. Linux-PCs mit adäquaten Massenspeichern. Weitere Details können von der Technik-Projektgruppe der UB Bielefeld bezogen werden.

[30] Ein offensichtliches Prinzip, welches für viele andere Gebiete in der Bibliothekswelt gilt.

[31] Elsevier benutzt die FAST-Software für einige seiner Produkte zur integrierten Suche (www.scirus.com und www.scopus.com), Thomsons »Web of Knowledge« basiert auf der Technologie »WebFeat Prism« (http://isiwebofknowledge.com/technology.html).

[32] Die Kerntechnologie für den Index ist »FAST Data Search«, zusätzliche Skripte für die Datenprozessierung und für die Benutzeroberflächen wurden mit Hilfe anderer Programmiersprachen und Open-Source-Technologie an der Universitätsbibliothek Bielefeld entwickelt.

[33] Summann, Friedrich; Lossau, Norbert: Search Engine Technoloy and Digital Libraries: Moving from Theory to Practice. In: D-Lib Magazine, September issue 2004. www.dlib.org/dlib/september04/lossau/09lossau.html. Eine PowerPoint-Präsentation, die auf einige Details eingeht, ist auf der Website der Digital Library Federation zugänglich. www.diglib.org/forums/Spring2004/summann0404.htm. – Die deutsche Übersetzung des Artikels erscheint in ZfBB 52 (2005) 1.

Der Verfasser

Dr. Norbert Lossau ist Direktor der Universitätsbibliothek Bielefeld, Universitätsstraße 25, 33615 Bielefeld, lossau@ub.uni-bielefeld.de